Resources for Machine Translation of the Macedonian Language

Katerina Zdravkova
Ss. Cyril and Methodius University in Skopje
67 PUBLICATIONS 232 CITATIONS

Resources for Machine Translation of the Macedonian Language

Stolić Miloš[1], Katerina Zdravkova[1]

[1] Institute of Informatics, Faculty of Natural Sciences and Mathematics
Arhimedova 5, 1000 Skopje, Macedonia
{Milos,Keti}@ii.edu.mk

Abstract. This paper focuses on creating new linguistic resources for the Macedonian language. It presents a new parallel corpus between Macedonian and Serbian language, build around the digitalized version of George Orwell's "1984", developed during the MULTEXT-EAST project. The original corpus is expanded with news articles from the Southeast European Times newspaper, published in public domain. The paper describes the retrieval, conversion, pre-processing, filtering and sentence-alignment of the corpus, then discusses and evaluates the alignment results.

Keywords: Natural Language Processing, Computational Linguistics, Bilingual Machine Translation, Statistical analysis, Language Resources

1 Introduction

Machine translation is one of the major fields of computational linguistics. Its usage and popularity are gradually increasing, with open source as well as proprietary technologies being constantly developed. Statistical models for machine translation are widely used today, mainly because vast amounts of parallel bi-texts are available online, and entire models can be learned automatically from them.

In order to efficiently use parallel texts, different correspondences between the target and source language must be identified. One way of identifying these correspondences is by means of parallel alignments. Depending on the type of parallel textual corpora, alignment can be done at document, paragraph, sentence, word or even at character level. It means that the alignment will identify the target document, paragraph, sentence, word and character corresponding to source ones.

The purpose of this paper is to identify existing linguistic resources for Macedonian language created during the MULTEXT-East project [1], and to present the creation process of sentence-aligned parallel bi-text corpora for Macedonian and Serbian language. The corpus is build from the online newspaper Southeast European Times [2], which is published in ten languages, so the described method can be used to generate any other language pair as well.

2 Lexical resources for the Macedonian language

2.1 The MULTEXT-East Project

MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages) [1] is a project for developing language resources for East European languages, which was a spin-off of the EU project MULTEXT. Although initiated a decade ago, resources for new languages are continuously being added to the dataset. At the moment, it provides resources for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Macedonian, Resian, Romanian, Russian, Slovene, Serbian and Persian language. Macedonian language has been included into Multext-East version 4. Its resources include morpho-syntactic descriptions (MSDs) of all 12 parts of speech (PoS), word-form lexica [3], fully annotated version of George Orwell's “1984” [4], as well as parallel bilingual and multilingual alignment at paragraph, and sentence level.

All translations of “1984” included in the MULTEXT-East dataset are sentence-aligned to the English version as a hub language. The size of the corpus is shown in Table 1.

Table 1. Size of the Orwell’s “1984“ corpus.

Language	Sentences	Total number of words	Unique word forms
Macedonian	6.630	98.764	15.689
Serbian	6.630	90.061	16.842

3 SETimes.com

The Southeast European Times (SET) is a Web site sponsored by the US Department of Defense [2]. It covers news and information from Southeastern Europe and it has been published since 2002. All information on the site is released in public domain and can be copied and distributed without permission. The articles are translated into 10 languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. As such, it is a potent and powerful resource for creating parallel bi-text corpora. This paper focuses on creating parallel corpora for Macedonian and Serbian language, but any other language pair can be similarly generated.

3.1 Technical details

The Southeast European Times web site is a standard dynamic html site. It does not contain any flash or similar multimedia content, so it can be easily downloaded using a web spider. The entire site weights 9.1 Gigabytes in approximately 400.000 files in total, or 980 Megabytes in 40.000 files per language. All html files were parsed and only the relevant text content was extracted. During the parsing, all html tags were removed.

For easier evaluation of the sentence alignment process, the content published in The Southeast European Times can be divided in four categories: articles, features, news briefs and roundups. Articles and features are generally longer texts, written in more than 20 sentences. News briefs are short news, mostly reported from other news agencies. Their length is up to 10 sentences. Finally, roundups are collections of very briefs news, usually a sentence or two, published several times during the day.

3.1.1 Language classification

Unfortunately, not all articles were actually translated in the appropriate language. Therefore, an N-gram language classifier was created for filtering out non-translated texts in all languages, as proposed in [5]. An N-gram is an N-character sequence of a longer string. The classifier works by creating per-language profiles. A profile is consisted of all N-grams found in a sample text in the corresponding language, sorted in reverse order by the number of occurrences.

A profile is also created for each unknown document that should be classified. It is compared to all previously determined per-language profiles, and a simple metric is used to determine the distance between them. This measure calculates how far out of place an N-gram in one profile is from its place in the other profile. For instance, if an N-gram is at rank 5 in one profile, i.e. it is the fifth most frequent N-gram, and at rank 7 in the other profile, the distance between them is 2. If an N-gram cannot be located in the reference profile, the distance takes some maximum value.

The sum of all distances is the distance measure for the document to the language. The document is then classified to the language that has the smallest distance measure.

After filtering out all identified texts, the archive was reduced to 26.606 files per language.

3.1.2 Pre-processing

The texts were typographically normalized - different quotation marks (Serbian « » ‘ and Macedonian „ “) were all translated into simple straight quotes " (U+0022 QUOTATION MARK). All files were concatenated in one large file and converted in XML using a several XSLT conversion scripts. Figure 1 shows the structure of the XML document for both languages. It is based on the recommendations of the Text Encoding Initiative, TEI P5 [6]. The number of word forms found in the corpus is shown in Table 2.

Table 2. Size of the Setimes corpus.

Language	Sentences	Total number of words	Unique word forms
Macedonian	266k	5.835k	111k
Serbian	265k	5.476k	135k

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE text SYSTEM "tei2.dtd">
<text id="mteo-mk." lang="mk">
 <body id="SETmk" lang="mk">
  <div type="part" n="1" id="SETmk.1">
   <p id="SETmk.1.1">
   <s id="SETmk.1.1.1">Романија продолжува
   мораториум во врска со меѓународните посвојувања
   на деца 29/10/2002 </s>
   <s id="SETmk.1.1.2">Под притисок на ЕУ Романија
   се согласи да ја продолжи забраната на
   меѓународни посвојувања на деца до 15 октомври
   менувајќи гипоранешните планови да ја отстрани
   овој месец </s>
```

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE text SYSTEM "tei2.dtd">
<text id="mteo-sr." lang="sr">
 <body id="SETsr" lang="sr">
  <div id="SETsr.1">
   <p id="SETsr.1.1">
   <s id="SETsr.1.1.1">Rumunija produzila
   moratorijum na međunarodna usvojenja dece
   29/10/2002 </s>
   <s id="SETsr.1.1.2">Pod pritiskom EU Rumunija je
   pristala da produzi moratorijum na međunarodna
   usvojenja dece do 15 novembra odustavsi od
   prethodne namere da zabranu ukine ovog meseca</s>
```

Fig. 1. Macedonian-Serbian parallel corpus sample.

3.2 Sentence alignment

The parallel corpus was automatically sentence-aligned using the HunAlign aligner. It is a language independent aligner developed under the Hunglish project [7]. Hunalign implements a hybrid algorithm of dictionary and length-based methods to align parallel bi-texts. It can work in two modes, with or without a starting dictionary. If no starting dictionary is given, HunAlign builds alignments based on Gale-Church's sentence length information [8]. Identical one-to-one translations found in this phase are extracted, and used as a starting dictionary. The alignment is then re-run, combining dictionary similarities with sentence length information.

Since Hunalign cannot read XML input, both file were converted to plain-text, tokenized and sentence-splitted using the Perl programs `tokenizer` and `split-sentences` written by Philipp Koehn, released with the Europarl corpus [9]. Document-level alignment was provided by inserting the `<p>` token after each article text.

The output of the alignment process is a three-column text file, where each line is consisted of a segment from the source text, followed by the corresponding segment from the destination text and confidence value for the segment. If one-to-many mappings are found, the second column contains zero or more sentences, separated by three consecutive tildes (~~~).

The Hunalign output was also converted to XML containing references to sentence IDs, as specified by the cesAlign DTD, an application of the Corpus Encoding Standard [10]. Figure 2 gives a Macedonian-Serbian alignment illustrating the syntax and types of alignment links: the first link encodes a 1-1 alignment, the second a 1-0, and the third an 1-2 alignment.

```
<link certainty="2.005722"
xtargets="SETmk.3.7.1;SETsr.3.5.1" id="SL13" />
<link certainty="0.73383" xtargets="SETmk.3.7.2; "
id="SL14" />
<link certainty="0.877679"
xtargets="SETmk.3.8.1;SETsr.3.6.1 SETsr.1.6.3"
id="SL17" />
```

Fig. 2. Example of bilingual alignment

3.3 Alignment evaluation

The success rate of the sentence-alignment was calculated using set of methods described in the ARCADE project [11]. We consider the *source text S* and its *translation text T* as two sets of segments $\{S_1, \ldots, S_n\}$ and $\{T_1, \ldots, T_n\}$. An *alignment A* is defined as a subset of the Cartesian product 2^S x 2^T. The triplet *(S, T, A)* is called a bi-text, and each of its elements a bisegment.

Let (S, T, A_r) be a bitext, and *A* an proposed alignment. The recall of the alignment *A* in reference to alignment A_r is the proportion of correct alignments in A with respect to the reference A_r:

$$recall = \frac{|A \cap A_r|}{|A_r|} \tag{1}$$

Precision stands for the proportion of bisegments in A that are correct with respect to the total of those proposed:

$$precision = \frac{|A \cap A_r|}{|A|} \quad (2)$$

F-measure combines the two previous metrics, and is defined as the harmonic mean of precision and recall (β=1):

$$F = 2 * \frac{recall * precision}{recall + precision} \quad (3)$$

The stated definitions of recall and precision do not account for partially correct bi-segments. The values range between 0 and 1 where a value close to 0 indicates a bad performance of the method while a value close to 1 indicates that the method performed very well. Table 3 shows the precision, recall and F-Measure calculated on a random sample from the sentence-aligned SETimes copus.

Table 3. Precision, Recall and F-Measure for the sentence-alignment task using Hunalign.

	Precision	Recall	F-Score
All alignments	0.940	0.955	0.947
1-1 alignments only	0.983	0.978	0.981

From the given results, it is clear that Hunalign performs significantly better on one-to-one alignments than on one-to-many alignments. One of the reasons for this is the actual content of the corpus. As shown in table 4, the results are much better on short documents (news briefs and roundups), where only one or a few lines need to be matched. Articles and features are sometimes not translated directly, and the authors add additional text, i.e. further explanation of a subject or their opinion, which is not present in the source text.

Table 4. Alignment quality of different parts of the corpus.

Category	Number of sentences	Precision
Articles	16.3k	0.930
Features	86.3k	0.900
News briefs	25.6k	0.990
Roundups	15.2k	0.980

The global alignment quality can obviously be improved by using a bi-language dictionary. It can be obtained either from the Hunalign output, or by word-aligning the sentence-aligned corpus, with tools like GIZA++ [12]. In either case, the automatically generated dictionary would have to be manually filtered for incorrect entries.

4 Conclusion and further work

This paper presents a parallel corpus between Macedonian and Serbian, based on public domain data. The corpus is comparable with other publicly available parallel corpora, in reference to the amount of data included and quality of alignment. The alignment accuracy can be improved by creating a large starting bi-language dictionary, as well as manually adding paragraph-level alignment tokens in texts with low results.

4.1 Part of speech tagging

The next step for this corpus would be part of speech tagging. The lexicon for Macedonian language created in the MULTEXT-East project contains 1.3 million word forms, while 52.3 thousand word forms found in the SETimes corpora are not present in it. However, most of these words are numbers, proper names or similar non-dictionary words. The statistical Trigrams'n'Tags part-of-speech tagger [13] tested in [3] performed with 98% accuracy, so similar results can be expected here as well.

4.2 Statistical machine translation

We believe that the amount of data found in the SETimes and “1984” corpuses is a solid starting point for training a statistical machine translation decoding system, such as Moses [14]. Due to the relatively small corpus, the decoder will probably give lower results compared to systems like Google Translate [15], but the accuracy will improve as the corpus is expanded.

Moses can be trained with phrase-based models and factored models. Phrase-based models map small chunks of text (phrase) between two bi-texts, with no use of linguistic information. On the other hand, factored models implemented in Moses expand phrase-based models with word-level annotation, like lemma, surface form, word class, morphology etc. This way, obvious SMT mistakes like linking a noun from the source language to a verb in the target language will be escaped. Depending on the accuracy of the POS-tagging process, a factored model can also be created from the SETimes corpus.

References

1. T. Erjavec: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Fourth International Conference on Language Resources and Evaluation, LREC-04, Paris, 2004
2. The Southeaster European Times, http://www.setimes.com
3. K. Zdravkova, A. Ivanovska, S. Džeroski, T. Erjavec: Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns. In Proceedings of IS

2005, the 8th International Multiconference on the Information Society, 11-17 October 2005, Ljubljana. pages. 195-198

4. V. Vojnovski, S. Dzeroski, T. Erjavec: Learning POS tagging from a tagged Macedonian text corpus. In Proceedings of IS 2005, the 8th International Multiconference on Information Society, 11-17 October 2005, Ljubljana, pages 199-202
5. W. Cavnar, J. Trenkle: N-Gram-Based Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161-175
6. TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/P5/
7. D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy: Parallel corpora for medium density languages. In Proceedings of the RANLP 2005 Conference, pages 590-596.
8. W. Gale, K. Church: A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Conference of the Association for Computational Linguistics, pages 177-184
9. P. Koehn: Europarl: A Parallel Corpus for Statistical Machine Translation. In Machine Translation Summit X, pages 79-86
10. Nancy Ide: Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In Proceedings of the First International Language Resources and Evaluation Conference, pages 463-470
11. V´ronis, J., Langlais, P.: Evaluation of parallel text alignment systems: The ARCADE project. In Vronis, J., ed.:Parallel text processing: Alignment and use of translation corpora, Kluwer Academic Publishers (2000), pages 369–388
12. F. Och, H. Ney: A Systematic Comparison of Various Statistical Alignment Models, In Computational Linguistics, volume 29, number 1, pages 19-51
13. T. Brants. TnT - a statistical part-of-speech tagger. In Proceedings of the Sixth Applied Natural Language Processing, Seattle, USA
14. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst: Moses: Open Source Toolkit for Statistical Machine Translation. In Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
15. Google Translate, http://translate.google.com